Mapping Between SNOMED RT and Clinical Terms Version 3: A Key Component of the SNOMED CT Development Process

Amy Y. Wang, MD[1], James W. Barrett, MB, ChB[2], Tim Bentley, MSc, BDS[2], David Markwell, MB, BS, MRCS, LRCP[3], Colin Price, MPhil, FRCS[2], Kent A. Spackman, MD, PhD[1,4], Michael Q. Stearns, MD[1]

[1]College of American Pathologists, Northfield, IL
[2]NHS Information Authority, Aston Cross, Birmingham, UK
[3]The Clinical Information Consultancy, Reading, UK
[4]Oregon Health Sciences University, Portland, OR

SNOMED® RT and Clinical Terms Version 3 are two large, controlled medical terminologies that are being merged to form a new work titled SNOMED® Clinical Terms (SNOMED® CT). One of the first steps in this process was to create maps between semantically equivalent and proximate concepts in the two terminologies. Same-as and is-a relationships were used to map the descriptions from one terminology to concepts in the other terminology. The objectives were to identify semantically equivalent concepts in the two terminologies, to find the most semantically proximate is-a relationships for non-equivalent concepts, and to evaluate the synonymy in the source terminologies. The results suggest that the rate of semantic overlap between descriptions in SNOMED RT and CTV3 is approximately 28%. This article discusses the methodology, issues, and findings of the description mapping process.

INTRODUCTION AND BACKGROUND

The Systematized Nomenclature of Medicine Reference Terminology (SNOMED RT), developed by the College of American Pathologists (CAP), and the United Kingdom's Clinical Terms Version 3 (Read Codes, abbreviated CTV3), are concept-based, multi-axial, controlled clinical terminologies.[1-3†] SNOMED RT and CTV3 both support enumerated and compositional functionality and have been shown in independent studies to perform favorably in a variety of medical settings and applications.[4,5] SNOMED RT is the most recent edition of SNOMED®, which has been developed over the past 35 years and historically has had strong representation of concepts used in acute care settings and pathology. With each new edition, SNOMED's content has expanded to represent a wider scope of terminology used in health care. The Read Codes were originally developed in 1985 for use in computerized patient records in primary care. CTV3, the latest version of the Read Codes, whose development began in 1992, was designed to support a wider range of healthcare settings.

SNOMED RT and CTV3 are both concept-based terminologies. A *concept* is a unit of thought that refers to a unique, clearly defined entity. An example is "Fundus of stomach." A *term* is a particular lexical string or expression that represents a concept. Terms are used in clinical information systems or other healthcare applications. In SNOMED RT, CTV3 and SNOMED CT, we use "description" to refer to terms that are linked to concepts in the core tables. This imparts a specific, nonambiguous meaning to each term. A single concept may have one or more associated descriptions. One description in each concept is designated the preferred name, and the others are called synonyms (Table 1). "Term" and "description" have often been used interchangeably in the past. However, the two are being distinguished because a term can be associated with different concepts in the clinical information systems depending on context, but a description is ideally non-ambiguous and always associated with a concept.

Concept code	Descriptions	Status
SNOMED RT	Cerebrovascular accident	Preferred name
D3-89550	CVA	Synonym
	Stroke	Synonym
CTV3	Stroke	Preferred name
X00D1	Cerebrovascular accident	Synonym
	CVA - Cerebrovascular accident	Synonym

Table 1. Concepts, descriptions, and synonyms.

In 1999, the CAP formed a strategic alliance with the NHS to merge SNOMED RT and CTV3 into a single terminology, SNOMED Clinical Terms (SNOMED

† SNOMED RT is a copyrighted work of the College of American Pathologists. Clinical Terms Version 3 was developed by the NHS and is a Crown copyright.

1067—5027/01/$5.00

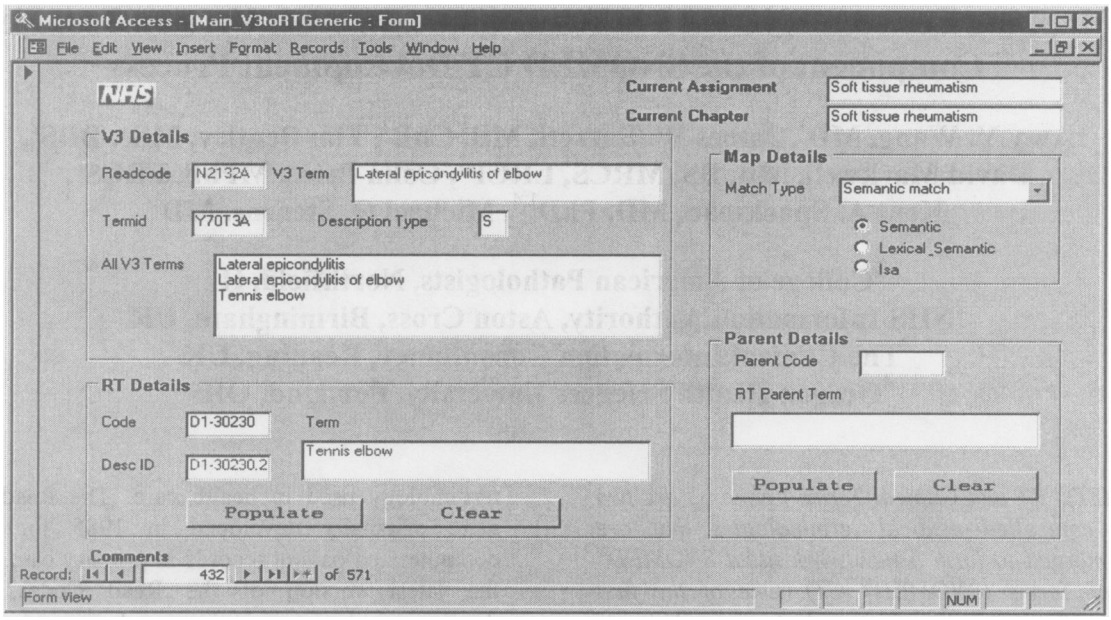

Figure 1. Description mapping tool used by US editors for mapping CTV3 descriptions to SNOMED RT concepts. "Lateral epicondylitis of elbow" is the CTV3 description under review. SNOMED RT does not have an exact lexical match. However, "Tennis elbow" is a synonym. A semantic match is made between "Lateral epicondylitis of elbow" from CTV3 and "Tennis elbow" from SNOMED RT.

CT). The ultimate goal of SNOMED CT is to increase the efficiency and reduce the cost of health care by improving the collection, storage, retrieval, and analysis of health care information in electronic patient records, clinical decision support, and research applications.[6]

The development of SNOMED CT requires the integration of over 320,000 concepts and 430,000 descriptions. SNOMED RT has over 120,000 active concepts and 160,000 active descriptions, and CTV3 has over 200,000 concepts and 270,000 descriptions. Ongoing development of the content of SNOMED RT and CTV3 and scheduled releases will continue throughout the merging process, necessitating updates of the data and increasing the complexity of the project. The structural differences between SNOMED RT and CTV3, the collaborative effort of many individuals and organizations, and the geographic and linguistic separation between the US and UK are also important factors in the process of content development for SNOMED CT.

An early step in the development of SNOMED CT was to create maps between descriptions in SNOMED RT and CTV3. Discussions about mapping medical terminologies together exist in the literature.[7-11] Most notably, multiple terminologies have been together to form the UMLS Metathesaurus, designed primarily to improve retrieval of information from electronic sources.[11-13] Other terminologies such as GALEN and LOINC have been mapped or are being mapped to SNOMED to facilitate integration of healthcare services or for research purposes.[14-16] However, the mapping between SNOMED RT and CTV3 is unique in its involvement of the developers of the source terminologies and the end goal of merging the two into one controlled medical terminology.

METHODS

Descriptions were mapped to concepts, rather than concepts to other concepts because each description in SNOMED RT and CTV3 is potentially an independent concept. This allowed editors to review the synonymy in each concept and identify potentially ambiguous concepts. This process also reduced the frequency of redundant concepts that may have existed in each source terminology.

The UK team developed a software tool in Microsoft Access for mapping descriptions, as shown in Figure 1. The tool contains a workflow module enabling a central administrator to assign editors sets of descriptions drawn from hierarchical sections of a terminology. This module ensures that the same sets of descriptions are never distributed more than once. The UK team also developed standalone browser software, which has been populated with data from SNOMED RT and CTV3. The browsers display concepts in a navigable link-based hierarchy and provide rapid and reliable retrieval of concepts and descriptions using code and keyword searches. The

standalone browsers were used in conjunction with the description mapping tool.

The UK editors mapped descriptions from SNOMED RT to concepts in CTV3. Editors in the US simultaneously mapped CTV3 descriptions to concepts in SNOMED RT. The US and UK mapping efforts differed somewhat. Both teams of editors mapped all descriptions in the findings, disorders, and procedures sections of both terminologies. This allowed for cross validation of work in the areas of greatest clinical focus. The UK editors also independently mapped descriptions in the anatomy, living organism, morphology, occupation, and substance (including drug) sections of SNOMED RT to concepts in CTV3.

Editors were instructed to evaluate a description based on its preferred name, the other descriptions associated with the same concept, and the meaning and usage of the concept, based on its supertypes, subtypes, attributes, and attribute values. Using the description mapping tool, US and UK editors assigned one of the following relationship types:

1. **Same-as** – There is a semantic match (i.e., synonymy) between the description in the source terminology and a concept in the target terminology. Editors were instructed to select the same-as designation only if the involved descriptions were pure synonyms with identical definitions. Same-as matches were divided into lexical-semantic and semantic types to aid in the process of removing lexically identical descriptions, but the two are considered together and referred to as same-as maps for the purposes of linking semantically equivalent concepts

    A. *Lexical-semantic* – The description from the source terminology has an exact string match with a description of the same meaning in the target terminology. "Renal disorder" from CTV3 and "Renal disorder" from SNOMED RT are semantically and lexically identical. Lexical equivalence alone was not sufficient for a match. For example, "Cold," connected with the SNOMED RT concept A-80210, a subtype of A-80190, "Temperature extreme," would not be mapped to "Cold," a synonym of "Common cold" which is associated with CTV3 concept XE0Xl.

    B. *Semantic* – The description being matched has the same meaning as a concept in the target terminology, but there is not an identical string match. For example, "Gastric disease," from CTV3 and "Disease of stomach" from SNOMED RT were treated as semantically but not lexically equivalent.

2. **Is-a** – When editors did not find a semantically equivalent concept in the target terminology, they were instructed to link the description to a semantically proximate supertype concept. For example, SNOMED RT has "Ulcer of pharynx," but CTV3 does not have a semantically equivalent concept. "Ulcer of pharynx" from SNOMED RT is-a "Disorder of pharynx" in CTV3.

3. **Unmappable** – There was no semantic match, and a supertype relationship could not be made. An editor may not have been able to clearly determine the meaning of a description.

When concepts with two or more non-synonymous descriptions were identified, the descriptions from the source terminology were separated by giving them same-as or is-a links to appropriate concepts in the target terminology. Semantically equivalent concepts from the two terminologies were identified and will be merged to form SNOMED CT concepts. A history mechanism will track the origin of each source concept and description from CTV3 and SNOMED RT.[17]

Conflicts occurred when editors disagreed on the semantic equivalence between descriptions and concepts in SNOMED RT and CTV3, or when an editor linked two descriptions in the same concept to different concepts in the target terminology. Descriptions from one terminology with no semantic matches were placed into the hierarchies of the other terminology through an is-a relationship. These relationships will be used to link semantically proximate concepts.

RESULTS

128,548 valid maps made from CTV3 to SNOMED RT by US editors were analyzed. This figure does not include the approximately 0.5% of reviewed CTV3 descriptions that were deemed unmappable. 29,016 (23%) were same-as maps and 99,532 (77%) were is-a maps. Of the 29,016 same-as maps, 19,763 (77%) were semantic only equivalents and 9,253 (23%) were lexical and semantic equivalents. We analyzed 140,873 valid maps from SNOMED RT to CTV3 made by the UK team. Excluded from this figure are the approximately 1.9% of reviewed SNOMED RT descriptions that the UK editors considered

unmappable. 47,731 (34%) were same-as maps and 93,142 (66%) were is-a maps. Of the same-as maps, 25,109 (53%) were semantic only matches and 22,622 (47%) were lexical and semantic matches. Of the 269,421 valid description maps from the US and UK that we reviewed, 76,747 (28%) were same-as maps and 192,674 (72%) were is-a maps.

Although editors mapped source terminology descriptions directly to target terminology descriptions, they were actually mapping to concepts in the target terminology. For example, if more than one CTV3 description (e.g., "Renal disease" and "Kidney disease") were given same-as maps to the same SNOMED RT concept (with its own descriptions including "Renal disease" and "Disease of kidney"), this is considered as one concept match, not two separate matches.

DISCUSSION

The US and UK editors found same-as description-to-concept maps 23 percent and 34 percent of the time, respectively. One reason for this discrepancy may be that the US and UK editors mapped different clinical areas. US editors focused their efforts on CTV3 findings, disorders, and procedures. In these areas, CTV3 has larger numbers of concepts and most likely has higher granularity than SNOMED RT (Table 2). In addition to these domains, the UK editors also mapped anatomy, living organisms, morphology, occupations, and substances from SNOMED RT. We found that the overall description-to-concept semantic equivalency rate was 28%. Somewhat higher numbers were expected, given the known structural and content-based similarities between SNOMED RT and CTV3.

Section	SNOMED RT	CTV3
Anatomy	10,980	9,325
Disorder	27,908	53,794
Finding/observation	6235	24,753
Living organism	20,721	5,543
Morphology	3,674	1,523
Occupation	1,865	3,133
Procedure	23,368	27,018
Substance/Drug	25,519	31,984

Table 2: Total concepts in select areas of SNOMED RT Version 1.0 and the September 2000 release of CTV3.

This project revealed significant differences between the content of SNOMED RT and CTV3. Although there is substantial overlap, the majority of descriptions and concepts are unique to each terminology. SNOMED RT has more anatomy, living organism and morphology concepts; CTV3 has more concepts describing diseases, findings, occupations, procedures, and substances. These differences may exist because of SNOMED's historical roots in pathology and CTV3's origins in primary care. We have shown that their differences in content will complement each other, with the merger resulting in a richer medical terminology.

Like SNOMED CT, the Unified Medical Language System (UMLS) Metathesaurus is a large, concept-based terminology.[18,19] Both merge concepts from source terminologies, utilize semantic relationships between concepts, and support multiple hierarchies. There are several key differences between the SNOMED CT and UMLS development processes.

The UMLS Metathesaurus contains over 800,000 concepts and 1.9 million concept names from numerous source terminologies, including SNOMED International (Version 3.5) and CTV3.[18,19] The merging of concepts into the UMLS Metathesaurus is accomplished by an algorithmic process, supplemented by manual review. The UMLS approach achieves its goal of linking a large number of sources together. In contrast, SNOMED CT, which will contain two merged terminologies, SNOMED RT and CTV3, is being developed by the editors who developed the source terminologies. Over thirty US and UK editors performed maps between SNOMED RT and CTV3 manually, with the major goal of preserving the original, intensional meanings of concepts. The SNOMED CT effort involves focused, manual review of each concept.

These differences reflect the intended purposes of SNOMED CT and the UMLS Metathesaurus. SNOMED CT is intended to represent patient data in electronic patient records. The UMLS Metathesaurus is designed to link a large number of source terminologies together to facilitate information retrieval. It would be an interesting exercise to compare the results of mapping SNOMED to CTV3 performed using the UMLS and SNOMED CT approaches.

Redundancy occurs when two concepts with identical meaning exist in a terminology. Conversely, ambiguity occurs when a concept contains two descriptions that are not synonymous and actually refer to different concepts. Mappings for findings, disorders and procedures were cross-validated. When editors in the US and UK agreed that all descriptions in two concepts were synonymous, they were merged into one concept. This is not considered a conflict, and the concept and its descriptions are added to the SNOMED CT concepts and descriptions tables.

Concept ambiguity may result from editing errors, variations in usage, or from "semantic drift," which is the alteration of meaning of a concept over time. Ambiguity is considered to be a type of conflict. All concepts that were identified as being potentially ambiguous during the mapping process were flagged for further review. These will be evaluated during the description mapping conflict resolution process.

Conflicts also occur when mapping efforts between the UK and US editors involving the same concept did not concur. The most common type was when editors were not in agreement as to whether two descriptions were synonymous. An editor may have made a deliberate decision or was not able to locate a matching concept in the target terminology. The former situation requires thorough review of each concept, and in some cases, consultation with a domain expert. The latter situation is readily resolved by an evaluation of the two mapping efforts. A tool specifically designed for resolving mapping conflicts has been developed, and the conflict resolution phase is under way at the time of this writing.

CONCLUSIONS

The description mapping process was an essential component of SNOMED CT development. We have demonstrated that it is possible to map two large controlled terminologies that have considerable differences in structure and content while preserving the exact semantic meanings of the original concepts. The full concept-modeling phase will provide further opportunities for review and quality assurance of SNOMED CT content.

Acknowledgments

The authors acknowledge the significant contributions from members of the SNOMED editorial board, SNOMED CT technical and content working groups, Kaiser Permanente CMT project, College of American Pathologists, NHS Information Authority, and the American Academy of Ophthalmology, including Diane Aschman, Peter Benton, Nick Booth, Phil Brown, James Campbell, Keith Campbell, Anne Casey, Robert Dolin, John Fedack, Paul Frosdick, Julia Kim, John Kilbourne, Debra Konicek, Karen Kudla, John Mason, Shelly Nash, Marcell Pooke, Marjorie Rallins, James Robb, David Robinson, Jeremiah Sable, Joshua Schraeder, and Corey Smith, and Monique van Berkum.

References

1. Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology for health care. Proc AMIA Symp 1997;640-4.
2. Spackman KA et al, eds. SNOMED RT. College of American Pathologists, Northfield, IL, 2000.
3. O'Neil M, Payne C, Read J. Read Codes Version 3: a user led terminology. Methods Inf Med 1995;34(1-2):187-92.
4. Campbell JR, Carpenter P, Sneiderman C, et al. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. JAMIA 1997;4:238-251.
5. Chute CG, Cohn SP, Campbell KE, et al. The content coverage of clinical classifications. JAMIA 1996;3:224-233.
6. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED Clinical Terms: Overview of the Development Process and Project Status. AMIA Annual Fall Symp 2001. In press.
7. Evans DA, Rothwell DJ, Monarch IA, Lefferts RG, Cote RA. Towards representations for medical concepts. Medical Decis Making;1991: 11 (supplement):S102-S108.
8. Kannry JL, Wright L, Shifman M, Silverstein S, Miller P. Portability issues for a structured clinical vocabulary: mapping from Yale to the Columbia medical entities dictionary. JAMIA 1996;3:66-78.
9. Wanger KM, Bentley TE, Maclean AR, Brown PJB, Price C. Early Version Read Codes: When is Non-Synonymy Important to the User? Proceedings of the Annual Conference of the Primary Health Care Specialist Group. Worcester, PHCSG, 1997; 69-79.
10. Zeng Q, Cimino JJ. Mapping medical vocabularies to the Unified Medical Language System. Proc AMIA Symp 1996;105-109.
11. Tuttle MS, Suarez-Munist ON, Olson NE, et al. Merging Terminologies. Medinfo 1995;162-166.
12. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabualaries. Proc AMIA Symp 1998;815-819.
13. Hole WT, Srinivasan S. Discovering missed synonymy in a large concept-oriented metathesaurus. Proc AMIA Symp 2000;354-358.
14. Rocha RA, Huff SM. Coupling vocabularies and data structures: lessons from LOINC. Proc AMIA Symp 1996;90-94.
15. Spackman KA. Integrating sources for a clinical reference terminology: experience linking SNOMED to LOINC and drug vocabularies. Medinfo. 1998;9 Pt 1:600-603.
16. Pole PM, Rector AL. Mapping the GALEN CORE Model to SNOMED-III: Initial Experiments. Proc AMIA Symp 1996;100-104.
17. SNOMED Clinical Terms technical specifications. College of American Pathologists, 2000. Available from: http://www.snomed.org.
18. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281-91.
19. UMLS Knowledge Sources. 12th ed. Bethesda: U.S. National Library of Medicine, 2001.
20. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Meth Inform Med 1998;37:394-403.